2018

Choosing cutoff values for correlated continuous diagnostic data to estimate sensitivity and specificity

Yingzhou Du
Iowa State University

Choosing cutoff values for correlated continuous diagnostic data to estimate sensitivity and specificity

by

Yingzhou Du

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Chong Wang, Co-major Professor
Huaiqing Wu, Co-major Professor
Jeffrey J. Zimmerman
Peng Liu
Daniel J. Nordman

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2018

DEDICATION

I would like to dedicate this dissertation to my maternal grandma and parents.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

ACKNOWLEDGMENTS

I want to take this opportunity to express my gratitude to those who helped me in completing my projects and writing this dissertation.

First, I want to show my deepest gratitude to my advisors, Dr. Chong Wang and Dr. Huaiqing Wu. Thank you Dr. Wang for your guidance and patience throughout this research. Thank you Dr. Wu for your suggestions on completing the projects and modifying the dissertation.

Next, I want to say thank you to Dr. Zimmerman for providing the detailed experimental design information.

I also want to express my thanks to Dr. Peng Liu and Dr. Dan Nordman for their willingness to be my committee members. Their good suggestion on my research is very helpful to me.

Finally, I want to show my gratitude to my family and all of my friends. It’s your love and encourage that help me advance day by day.

ABSTRACT

This dissertation includes two projects and an R package for the 1st project, focusing on choosing cutoff values to estimate the sensitivity and specificity for continuous diagnostic tests. A continuous diagnostic test needs to be dichotomized to generate positive and negative test outcomes by choosing a cutoff value. The choice of the cutoff value depends on the estimates of the sensitivity and specificity of the dichotomized test with this cutoff. There are two challenges during this process: 1) a typical experiment to validate a new diagnostic test usually involves multiple observations from the same subjects, resulting in correlated data values. This correlation increases the complexity of calculating the confidence intervals of the sensitivity and specificity given the true statuses. 2) In many diagnostic trials, the true statuses are unknown, which make it difficult even to calculate the point estimates of the sensitivity and specificity. In the 1st project (Chapter 2), we propose a method to calculate the confidence intervals of sensitivity and specificity with the true statuses of subjects given based on a parametric model for the correlated continuous diagnostic test data. In the second project (Chapter 3), we focus on the challenge of unknown statuses. Due to the lack of the statuses, only the model-based method can be used to estimate the sensitivity and specificity. However, the estimations of model parameters are difficult because of the unknown statuses. We propose a Bayesian model with latent variables to model the unknown statuses. In Chapter 4, I create an R package for the 1st project for future uses.

CHAPTER 1. GENERAL INTRODUCTION

1.1 Background

A receiver operating characteristic (ROC) curve is a useful tool to assess and compare binary classification tests [Delong 1988, Hanley 1982, and Metz 1978]. Sensitivity and specificity are building bricks of ROC curves. Sensitivity, also known as the true positive rate, is defined as the proportion of positive tests for the diseased group, which is written as $\Pr(T+|D)$, where D denotes the diseased group and $T+$ denotes the positive tests. Similarly, specificity, also known as the true negative rate, is defined as $\Pr(T-|D^c)$, where D^c denotes the non-diseased group and $T-$ denotes the negative tests. Nowadays, the ROC curve is widely used in continuous diagnostic tests. For continuous data, positive and negative test results are separated by the select cutoff value. However, there are two challenges during the process of estimating the sensitivity and specificity: 1) A typical experiment to validate a new diagnostic test usually involves multiple observations from the same subjects, resulting in correlated data values. The correlation influences the variances of the estimated sensitivity and specificity and thus the confidence intervals (CIs). 2) In many clinical diagnostic trials, the true statuses of patients are unknown. The unknown statuses will i) make the model-based method to be the only choice to estimate the sensitivity and specificity ii) increase the complexity to analyze the data. The construction of proper statistical methods for either of these two challenges is crucial for estimating the sensitivity and specificity and choosing the cutoffs.

In this dissertation, we propose a model based on the normal distribution to analyze the correlated data and then calculate the CIs of sensitivity and specificity based on the model parameters for the 1st challenge; we also propose a model with a latent variable to model the

data with unknown statuses for the 2nd challenge. Due to the complexity of the model, we use the Bayesian method together with the Markov chain Monte Carlo (MCMC) to make inference for the model parameters.

1.2 Dissertation Organization

This dissertation consists of three main chapters (Chapters 2, 3, and 4). In the 1st project (Chapter 2), we propose a model based on the normal distribution for the correlated and continuous data. The variances of the estimated sensitivity and specificity are derived as functions of model parameters. The logit transformation and the corresponding back transformation are applied to avoid the case that the CIs exceed the range of [0, 1]. We apply both the newly-proposed method and the existing methods to simulated data sets to compare the performances of different methods. In the 2nd project (Chapter 3), we model the unknown statuses as a latent variable in the proposed normal-distribution based model. The Bayesian method combined with MCMC is used to get the samplings of the posterior distribution of each model parameter. The model parameters are estimated from the samplings of the corresponding posterior distributions. Given a desired value of sensitivity or specificity, the cutoff is estimated by inversing the model-based estimate of sensitivity or specificity. We perform the simulation studies to see if our method can recover the true model parameters and the theoretical value of the cutoff for the desired sensitivity or specificity. In Chapter 4, I build an R package for the 1st project for future use. The manual of the R project is the main component of Chapter 4. Finally, I make a general discussion and conclusion in Chapter 5.

CHAPTER 2. CHOOSING CUTOFF VALUES FOR CORRELATED CONTINUOUS DIAGNOSTIC DATA: ADJUSTMENT FOR CONFIDENCE INTERVAL ESTIMATIONS OF SENSITIVITY AND SPECIFICITY

Abstract: A continuous diagnostic test needs to be dichotomized to generate positive and negative test outcomes by choosing a cutoff value. The choice of the cutoff value depends on the estimates of the sensitivity and specificity of the dichotomized test with this cutoff. A typical experiment to validate a new diagnostic test usually involves multiple observations from the same subjects, resulting in correlated data values. Currently, methods to calculate interval estimates of sensitivity and specificity are available for binary data. In this paper, we propose a method to calculate the confidence intervals of sensitivity and specificity based on a parametric model for the correlated continuous diagnostic test data. Simulation studies show that our proposed method outperforms the current methods for binary correlated data in terms of the coverage probability and calculating efficiency. In addition, we provide an application to choosing cutoff values for the influenza A virus test in swine.

2.1 Introduction

As discussed in section 1.1, sensitivity and specificity are important statistical measures for binary tests. In many biological clinical tests, the enzyme-linked immunosorbent assay (ELISA) or the polymerase chain reaction (PCR) is used to collect data from multiple subjects. The raw data for these tests are usually continuous and correlated. For example, a typical veterinary test often collects data at different times for each of a group of subjects (usually swine or rats). Data collected from the same subjects are correlated while data from different subjects tend to be uncorrelated. For such data, a cutoff point is selected to separate positive and negative test results and further to evaluate sensitivity and specificity. Large values of

sensitivity and specificity are favorable. However, the magnitudes of sensitivity and specificity change along opposite directions as the cutoff changes. Therefore, a trade-off between sensitivity and specificity needs to be made by selecting a proper cutoff. The intra-subject correlation usually does not affect the point estimates of sensitivity and specificity. However, the interval estimates are problematic if the correlation is ignored since correlations are usually positive and thus inflate the variances. Currently, many methods are available to calculate confidence intervals (CIs) for sensitivity and specificity for binary data. Methods that ignore the correlations, such as the simple asymptotic method, the Clopper-Pearson method, the Wilson method, and the Agresti-Coull method, fail to account for the inflated variances and thus result in lower coverage probabilities [Agresti 1998, Clopper 1934, NEWCOMBE 1998, and Wilson 1927]. Alternative methods taking the binary correlations into consideration are available in the literature. Some non-parametric methods modify the simple asymptotic method by either adjusting the standard error with a ratio (SER) factor or a variance influence factor (VIF) or re-estimating the variance with a between-group variance (BGV) method [Fleiss 2003, Rao 1992, and Williams 2000]. There are also some parametric methods, such as the generalized linear mixed model (GLMM) and the generalized estimating equation (GEE), which work for correlated binary data. The GLMM assumes the binomial distribution for the data. Sensitivity and specificity, which are essentially binary proportions, are usually linked by a logistic function and estimated by conditioning on the random subjects [Mutsvari 2010]. The overall estimates of sensitivity and specificity are the medians of these subject-specific sensitivities and specificities, respectively. The GEE, which is used to implement a marginal model, requires a working correlation matrix for the correlations within the same subjects

[Smith 1992 and Zeger 1986]. The estimated sensitivity (and specificity) is the weighted average across all observations.

All these methods developed for binary data theoretically apply to the continuous data since they can be dichotomized to binary data by a selected cutoff. However, there are two challenges for these methods: 1) If the true sensitivity or specificity are large, almost all tests might be positive or negative especially for small- to medium-sized data. This will result in CIs with width close to 0 for nonparametric methods and convergence problems for model-based methods. 2) If different cutoffs are chosen, different converted binary data are obtained for different cutoffs. This will increase the burden of calculation if binary model-based methods are used. In this study, considering that most continuous clinical test data are normally distributed or normally distributed after transformation, we propose a new method based on the normal distribution to calculate the CIs of sensitivity and specificity for correlated and normally distributed observations.

2.2 Method

2.2.1 Model

As discussed in the introduction section, in a typical biological clinical trial, multiple observations are often collected from each of many subjects. Since observations within the same subject tend to be correlated, a linear mixed model was assumed for data analysis:

$$Y_{ij} = \mu + \gamma_i + \tau s_{ij} + \epsilon_{ij}$$

where Y_{ij} is the j^{th} observation for the i^{th} subject, $i = 1,2,\dots,P, j = 1,2,\dots,n_i$; μ is the overall mean for the non-diseased (or healthy) group; $\gamma_i' s \overset{iid}{\sim} N(0, \sigma_\gamma^2)$, where γ_i denotes the random effect of the i^{th} subject; τ, the mean difference between the diseased group and the

healthy group, is a fixed effect; $s_{ij} = I(Y_{ij}$ is measured from the diseased group) denotes the status of the j^{th} observation for the i^{th}subject. Here $I(\cdot)$ is the indicator function and defined as $I(A) = 1$ if and only if A is true. The $\epsilon_{ij} \overset{\text{indep}}{\sim} N(0, \sigma_{\text{e}_s}^2)$, where ϵ_{ij} denotes the error term associated with the j^{th} observation for the i^{th} subject and $s \triangleq s_{ij}$. Based on this model, we obtain the following means and covariances,

$$E(Y_{ij}) = \begin{cases} \mu \triangleq \mu_0 & \text{healthy } (s_{ij} = 0) \\ \mu + \tau \triangleq \mu_1 & \text{diseased } (s_{ij} = 1) \end{cases}$$

$$\text{Cov}(Y_{ij}, Y_{i'j'},) = \sigma_\gamma^2 I(i = i') + \sigma_{\text{e}_s}^2 I(i = i') I(j = j')$$

$$= \begin{cases} 0 & \text{if } i \neq i' \\ \sigma_\gamma^2 + \sigma_{\text{e}_s}^2 \triangleq \sigma_s^2 = \text{Var}(Y_{ij}) & \text{if } i = i',\ j = j' \\ \sigma_\gamma^2 & \text{otherwise} \end{cases} \tag{2.1}$$

Therefore, if $(i, j) \neq (i', j')$,

$$\begin{pmatrix} Y_{ij} \\ Y_{i'j'} \end{pmatrix} \sim \text{MVN}_2 \left(\begin{pmatrix} \mu_s \\ \mu_{s'} \end{pmatrix}, \begin{pmatrix} \sigma_\gamma^2 + \sigma_{\text{e}_s}^2 & \sigma_\gamma^2 I(i = i') \\ \sigma_\gamma^2 I(i = i') & \sigma_\gamma^2 + \sigma_{\text{e}_{s'}}^2 \end{pmatrix} \right), \tag{2.2}$$

where $s' \triangleq s_{i'j'}$.

2.2.2 Estimates of sensitivity and specificity and their variances

2.2.2.1 Point estimates of sensitivity and specificity

In the following discussion, we assume the mean response of the diseased group to be larger unless specified. Sensitivity and specificity are abbreviated as Sen and Spe in mathematical expressions. For a specific cutoff c, the empirical estimates of sensitivity and specificity are defined as

$$\widehat{\text{Sen}} = \sum_{i=1}^{P}\sum_{j=1}^{n_i} I(Y_{ij} \geq c)I(s_{ij} = 1)/N_d = \sum_{i=1}^{P}\sum_{j_d=1}^{d_i} I(Y_{ij_d} \geq c)/N_d \qquad (2.3)$$

$$\widehat{\text{Spe}} = \sum_{i=1}^{P}\sum_{j=1}^{n_i} I(Y_{ij} < c)I(s_{ij} = 0)/N_h = \sum_{i=1}^{P}\sum_{j_h=1}^{h_i} I(Y_{ij_h} < c)/N_h, \qquad (2.4)$$

where $N_d, d_i,$ and j_d are the total number of observations, the number of observations within subject i, and the index for the diseased group. And N_h, $h_i = n_i - d_i$, and j_h are the corresponding terms for the healthy group.

An alternative model-based estimates of sensitivity and specificity, denoted by "*", are defined as

$$\widehat{\text{Sen}}^* = \widehat{\Pr}(T+|D) = \widehat{\Pr}\big(Y_{ij} > c|D\big) = 1 - \Phi\left(\frac{c-\hat{\mu}_1}{\hat{\sigma}_1}\right), \qquad (2.5)$$

where $\widehat{\Pr}(\cdot)$ denotes the probability calculated with respect to an estimated distribution for the diseased group, and $\Phi(\cdot)$ is the standard normal cumulative distribution function. Similarly,

$$\widehat{\text{Spe}}^* = \Phi\left(\frac{c-\hat{\mu}_0}{\hat{\sigma}_0}\right). \qquad (2.6)$$

The expected values of $\widehat{\text{Sen}}$ and $\widehat{\text{Spe}}$ are

$$E\big(\widehat{\text{Sen}}\big) = \sum_{i=1}^{P}\sum_{j_d=1}^{d_i} E(I(Y_{ij_d} \geq c))/N_d = \Pr(Y_{ij_d} \geq c) = \Pr(T+|D) = \text{Sen},$$

$$E\big(\widehat{\text{Spe}}\big) = \sum_{i=1}^{P}\sum_{j_h=1}^{h_i} E(I(Y_{ij_h} < c))/N_h = \Pr\big(Y_{ij_h} < c\big) = \Pr(T-|H) = \text{Spe}.$$

Therefore, $\widehat{\text{Sen}}$ and $\widehat{\text{Spe}}$ are unbiased estimators of sensitivity and specificity, no matter what the correlation is.

2.2.2.2 Variances of $\widehat{\text{Sen}}$ and $\widehat{\text{Spe}}$

The derivations of the variances for $\widehat{\text{Sen}}$ and $\widehat{\text{Spe}}$ are shown below:

$$\text{Var}\left(\widehat{\text{Sen}}\right) = \text{Var}\left(\sum_{i=1}^{P}\sum_{j_d=1}^{d_i} I(Y_{ij_d} \geq c)/N_d\right)$$

$$= \text{Var}\left(1 - \sum_{i=1}^{P}\sum_{j_d=1}^{d_i} I(Y_{ij_d} \geq c)/N_d\right)$$

$$= \text{Var}\left(\sum_{i=1}^{P}\sum_{j_d=1}^{d_i} I(Y_{ij_d} < c)/N_d\right)$$

$$= \frac{1}{N_d^2}\text{Var}\left(\sum_{i=1}^{P}\sum_{j_d=1}^{d_i} I(Y_{ij_d} < c)\right)$$

$$= \frac{1}{N_d^2}\text{Cov}\left(\sum_{i=1}^{P}\sum_{j_d=1}^{d_i} I(Y_{ij_d} < c), \sum_{i'=1}^{P}\sum_{j'_d=1}^{d_i} I(Y_{i'j'_d} < c)\right)$$

$$= \frac{1}{N_d^2}\left\{E\left(\left[\sum_{i=1}^{P}\sum_{j_d=1}^{d_i} I(Y_{ij_d} < c)\right]\left[\sum_{i'=1}^{P}\sum_{j'_d=1}^{d_i} I(Y_{i'j'_d} < c)\right]\right) - \left(E\left[\sum_{i=1}^{P}\sum_{j_d=1}^{d_i} I(Y_{ij_d} < c)\right]\right)^2\right\}$$

$$= \frac{1}{N_d^2}\left\{E\left[\sum_{i=1}^{P}\sum_{j_d=1}^{d_i} I(Y_{ij_d} < c)\right]\right\}$$

$$+ \frac{1}{N_d^2}\left\{E\left[\sum_{i,j_d,j'_d \neq j_d} I\left(Y_{ij_d} < c\right)I\left(Y_{ij'_d} < c\right) + \sum_{i,i' \neq i,j_d,j'_d} I\left(Y_{ij_d} < c\right)I\left(Y_{i'j'_d} < c\right)\right]\right\}$$

$$- \frac{1}{N_d^2}\left(N_d\left(1 - \text{Sen}\right)\right)^2$$

$$= \frac{1}{N_d^2}\left\{N_d(1-\text{Sen}) + \left(\sum_i [d_i(d_i-1)]\right)\Pr\left(Y_{ij_d} < c, Y_{ij_d'} < c\right)\right.$$

$$\left. + \left(N_d^2 - N_d - \sum_i [d_i(d_i-1)]\right)(1-\text{Sen})^2\right\} - (1-\text{Sen})^2$$

$$= \frac{1}{N_d}(1-\text{Sen})$$

$$+ \frac{\sum_i [d_i(d_i-1)]}{N_d^2}\Pr\left(Y_{ij_d} - \mu_1 < c - \mu_1, Y_{ij_d'} - \mu_1\right.$$

$$\left. < c - \mu_1\right) - \frac{N_d + \sum_i [d_i(d_i-1)]}{N_d^2}(1-\text{Sen})^2.$$

Since $\begin{pmatrix} Y_{ij_d} \\ Y_{ij_d'} \end{pmatrix} \sim MVN_2\left(\begin{pmatrix} \mu_1 \\ \mu_1 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_\gamma^2 \\ \sigma_\gamma^2 & \sigma_1^2 \end{pmatrix}\right)$, we have $\begin{pmatrix} Y_{ij_d} - \mu_1 \\ Y_{ij_d'} - \mu_1 \end{pmatrix} \sim MVN_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_\gamma^2 \\ \sigma_\gamma^2 & \sigma_1^2 \end{pmatrix}\right)$.

Define $\Sigma_s = \begin{pmatrix} \sigma_s^2 & \sigma_\gamma^2 \\ \sigma_\gamma^2 & \sigma_s^2 \end{pmatrix}$ and $c_s = c - \mu_s$, and let $F_s(x, y)$ be the cumulative distribution function of $MVN_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_s\right), s = 0$ or 1. Then

$$\text{Var}(\widehat{\text{Sen}}) = \frac{1}{N_d}(1-\text{Sen}) + \frac{\sum_i [d_i(d_i-1)]}{N_d^2}F_1(c_1, c_1) - \frac{N_d + \sum_i [d_i(d_i-1)]}{N_d^2}(1-\text{Sen})^2. \quad (2.7)$$

Similarly,

$$\text{Var}(\widehat{\text{Spe}}) = \frac{1}{N_h}(\text{Spe}) + \frac{\sum_i [h_i(h_i-1)]}{N_h^2}F_0(c_0, c_0) - \frac{N_h + \sum_i [h_i(h_i-1)]}{N_h^2}(\text{Spe})^2. \quad (2.8)$$

The derivation above is based on the assumption of larger means for the diseased group. If, on the contrary, the healthy group possesses larger means, then formulas (2.7) and (2.8) would be modified as

$$\mathrm{Var}\left(\widehat{\mathrm{Sen}}\right) = \frac{1}{N_d}(\mathrm{Sen}) + \frac{\sum_i [d_i(d_i-1)]}{N_d^2} F_1(c_1, c_1) - \frac{N_d + \sum_i [d_i(d_i-1)]}{N_d^2}(\mathrm{Sen})^2, \tag{2.7a}$$

$$\begin{aligned}\mathrm{Var}\left(\widehat{\mathrm{Spe}}\right) = {} & \frac{1}{N_h}(1-\mathrm{Spe}) \\ & + \frac{\sum_i [h_i(h_i-1)]}{N_h^2} F_0(c_0, c_0) - \frac{N_h + \sum_i [h_i(h_i-1)]}{N_h^2}(1-\mathrm{Spe})^2.\end{aligned} \tag{2.8a}$$

2.2.2.3 Interval estimates of sensitivity and specificity

The most straightforward method to calculate the 95% CIs of sensitivity and specificity is

$$\widehat{\mathrm{Sen}} \pm z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\widehat{\mathrm{Sen}})} \tag{2.9}$$

and

$$\widehat{\mathrm{Spe}} \pm z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\widehat{\mathrm{Spe}})}, \tag{2.10}$$

where $\widehat{\mathrm{Var}}\left(\widehat{\mathrm{Sen}}\right)$ and $\widehat{\mathrm{Var}}(\widehat{\mathrm{Spe}})$, the estimates of $\mathrm{Var}\left(\widehat{\mathrm{Sen}}\right)$ and $\mathrm{Var}\left(\widehat{\mathrm{Spe}}\right)$, are calculated by replacing the unknown parameters with their corresponding estimates; $z_{1-\alpha/2}$ is the $100\left(1-\frac{\alpha}{2}\right)$% percentile of the standard normal distribution.

However, these interval estimates of sensitivity and specificity might exceed the range of a probability, i.e., [0, 1]. To avoid this problem, we use the idea of transformation for constructing the CIs: 1) apply the logit-transformation to the estimates of sensitivity and

specificity, and 2) construct CIs for the logit-transformed $\widehat{\text{Sen}}$ and $\widehat{\text{Spe}}$. The variances of the transformed $\widehat{\text{Sen}}$ and $\widehat{\text{Spe}}$ are derived by the delta method, as shown below

$$\text{Var}\left(\text{logit}\left(\widehat{\text{Sen}}\right)\right) = \frac{\text{Var}\left(\widehat{\text{Sen}}\right)}{[\text{Sen}(1-\text{Sen})]^2}$$

$$\text{Var}\left(\text{logit}\left(\widehat{\text{Spe}}\right)\right) = \frac{\text{Var}\left(\widehat{\text{Spe}}\right)}{[\text{Spe}(1-\text{Spe})]^2}.$$

The 95% CIs for logit(Sen) and logit(Spe) are

$$\text{logit}\left(\widetilde{\text{Sen}}\right) \pm z_{1-\alpha/2}\sqrt{\frac{\widehat{\text{Var}}^*\left(\widehat{\text{Sen}}\right)}{\left[\left(\widehat{\text{Sen}}^*\right)\left(1-\widehat{\text{Sen}}^*\right)\right]^2}} \tag{2.11}$$

and

$$\text{logit}\left(\widetilde{\text{Spe}}\right) \pm z_{1-\alpha/2}\sqrt{\frac{\widehat{\text{Var}}^*\left(\widehat{\text{Spe}}\right)}{\left[\left(\widehat{\text{Spe}}^*\right)\left(1-\widehat{\text{Spe}}^*\right)\right]^2}}, \tag{2.12}$$

where $\widehat{\text{Var}}^*\left(\widehat{\text{Sen}}\right)$ and $\widehat{\text{Var}}^*\left(\widehat{\text{Spe}}\right)$, similar to $\widehat{\text{Var}}\left(\widehat{\text{Sen}}\right)$ and $\widehat{\text{Var}}\left(\widehat{\text{Spe}}\right)$, are another estimates of $\text{Var}\left(\widehat{\text{Sen}}\right)$ and $\text{Var}\left(\widehat{\text{Spe}}\right)$. The only difference is that $\widehat{\text{Var}}^*\left(\widehat{\text{Sen}}\right)$ and $\widehat{\text{Var}}^*\left(\widehat{\text{Spe}}\right)$ use the model-based estimates of sensitivity and specificity for calculation to avoid negative variances or infinite variances. For the point-estimates part of the CIs, the empirical estimates are preferred. However, as discussed by [Brown 2001], when very few (usually ≤ 5) observations are either below or above the cutoffs, the empirical point estimate of the proportion is unsuitable for calculating the CIs. In this case, the model-based point estimates are used. Therefore, in the point-estimates part of the CIs,

$$\widetilde{\text{Sen}} \triangleq \begin{cases} \widehat{\text{Sen}} & \text{if } 5 < \text{Number of } (T+|D) < N_d - 5 \\ \widehat{\text{Sen}}^* & \text{otherwise} \end{cases}$$

and

$$\widetilde{\text{Spe}} \triangleq \begin{cases} \widehat{\text{Spe}} & \text{if } 5 < \text{Number of } (T - |H) < N_h - 5 \\ \widehat{\text{Spe}}^* & \text{otherwise} \end{cases}.$$

Finally, the interval estimates of sensitivity and specificity are obtained by back-transforming the CIs of the transformed sensitivity and specificity.

2.3 Simulations

The purpose of the simulation study is to evaluate if CIs calculated by the proposed method have reasonable coverage probabilities.

2.3.1 Parameters

We chose 30 subjects for the simulation, with 21 observations from each of the subjects. Among these subjects, P_1 (= 5) subjects were always healthy and data from these subjects were considered from the healthy group. For the other P_2 (= 25) subjects, n_{21} (= 3) observations were collected from each subject when the subjects were healthy while the other n_{22} (= 18) were collected when the subjects were infected. The parameters for the simulation were chosen as $\mu_0 = 0;\ \mu_1 = -0.8;\ \sigma_\gamma^2 = 0.01, 0.1, 0.5; \sigma_{e_0}^2 = 0.02; \sigma_{e_1}^2 = 0.1$. For each combination of these parameters, 10,000 datasets were simulated. Sensitivity and specificity were calculated w.r.t cutoff = $(-0.8, -0.6, -0.4, -0.2, 0)$. The simulated data and the following real data were both analyzed by a linear mixed model package in R (R: nlme).

2.3.2 Results and discussion

2.3.2.1 Estimates of standard errors of $\widehat{\text{Sen}}$ and $\widehat{\text{Spe}}$

Since in the simulated data the non-diseased group possessed a larger mean, formulas (2.7 a) and (2.8 a), instead of formulas (2.7) and (2.8), were used to calculate the standard errors of $\widehat{\text{Sen}}$ and $\widehat{\text{Spe}}$. The averaged standard errors were calculated by averaging the standard

errors over 10,000 simulations. The standard errors were also estimated by the simple asymptotic method and averaged over 10,000 simulations. The theoretical standard errors were calculated by plugging in the true parameters into (2.7 a) and (2.8 a). The plots of all the standard errors are shown in Figure 2.1. Note that the standard errors estimated by the proposed method are very close to the true values while the standard errors estimated by the simple asymptotic method are significantly below the true values, especially for large covariances.

2.3.2.2 Coverage Probabilities

From each combination of the selected cutoffs and the other simulation parameters, 10,000 95% CIs of sensitivity and specificity were calculated from 10,000 simulations by following the procedure discussed in Section 2.2. The coverage probability (CP) is the probability of the calculated CIs covering the true value. The true values of sensitivity and specificity are calculable by replacing estimates of model parameters with the true parameters in formulas (2.5) and (2.6). The CPs of the CIs calculated by the proposed method and the existing methods mentioned in the introduction are plotted in Figure 2.2. Note that several methods can be used to calculate CIs with correlations ignored. We use the Clopper-Pearson method because it is conservative.

Figure 2.2 shows that the proposed method (red solid lines with hollow circular symbols) outperforms all of the existing methods. The CP calculated using the method with correlations ignored is significantly below 95% because it fails to account for the inflated variation introduced by the correlation. The GLMM doesn't perform well for data with a large enough intra-subject correlation (covariance ≥ 0.1). The GEE, on the contrary, performs well except when the sensitivity or specificity is extremely close to 1. However, both of the two

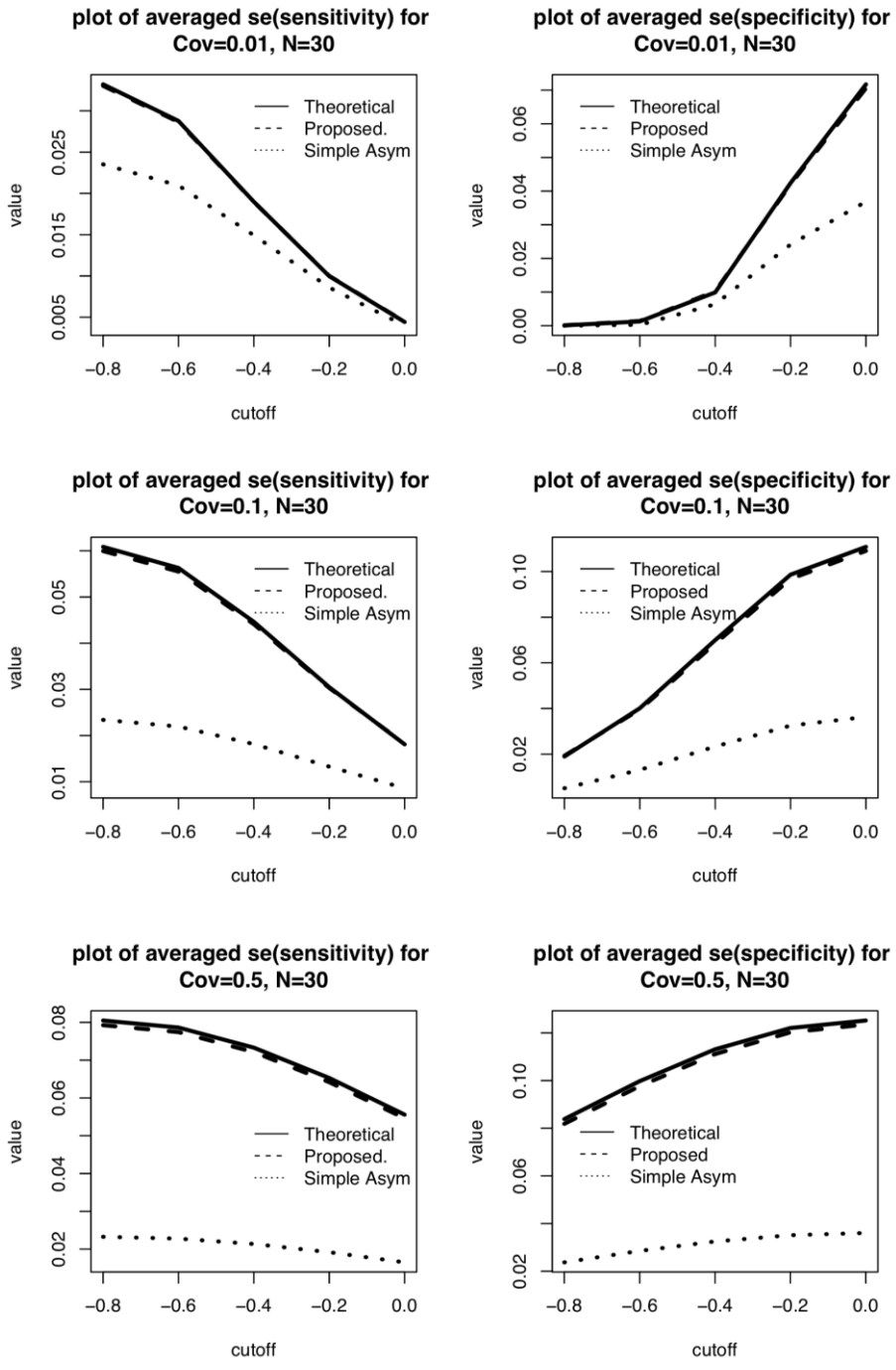

Figure 2.1. Plots of the averaged estimates of standard errors and theoretical standard errors: solid lines denote theoretical values; dashed lines denote the averaged estimates from the proposed method; dotted lines denote averaged estimates from the simple asymptotic method.

current model-based methods suffer from convergence problems and many CPs showed in Figure 2.2 are calculated based on part of the simulated datasets. Table 2.1 shows the number of converged analyses out of the total number of simulated data. From Table 2.1, we observe that both extreme values of sensitivity or specificity and large correlations (given similar values of sensitivity and specificity) are detrimental to the convergence for the GEE method. And there are more non-convergences for the GLMM. When the covariance is 0.01 and the cutoff is -0.8, there are only five converged analyses for the calculation of the CP of specificity. This is because with these parameter values, the specificity is 0.999981 and the probability that all of the 180 simulated clinical test results in the healthy group are negative is 99.966%.

As for the non-parametric method, when the numbers of observations from each subject are not too small, just like the simulated data in the diseased group, the three non-parametric methods, VIF, SER and BGV, work well and have similar CPs. However, as the magnitude of sensitivity or specificity gets close to 1 or the covariance increases, the CPs decrease. On the other hand, when most of subjects don't have many observation, like the healthy group, the CPs of VIF and SER decrease significantly from the nominal value, especially when covariance is large or the sensitivity or the specificity is extremely large in magnitude. The BGV is very conservative and the CPs tend to be larger than the nominal value.

The CPs calculated with the proposed method are very close to the nominal value 95% and only slightly positively biased when sensitivity or specificity is above ~90%. Besides the reasonable CPs, another advantage of the proposed method over current model-based methods is that a new cutoff does not create a new dataset, which means that we need not re-analyze the data for a new cutoff value. This is important when performing large volumes of simulations with different cutoff values.

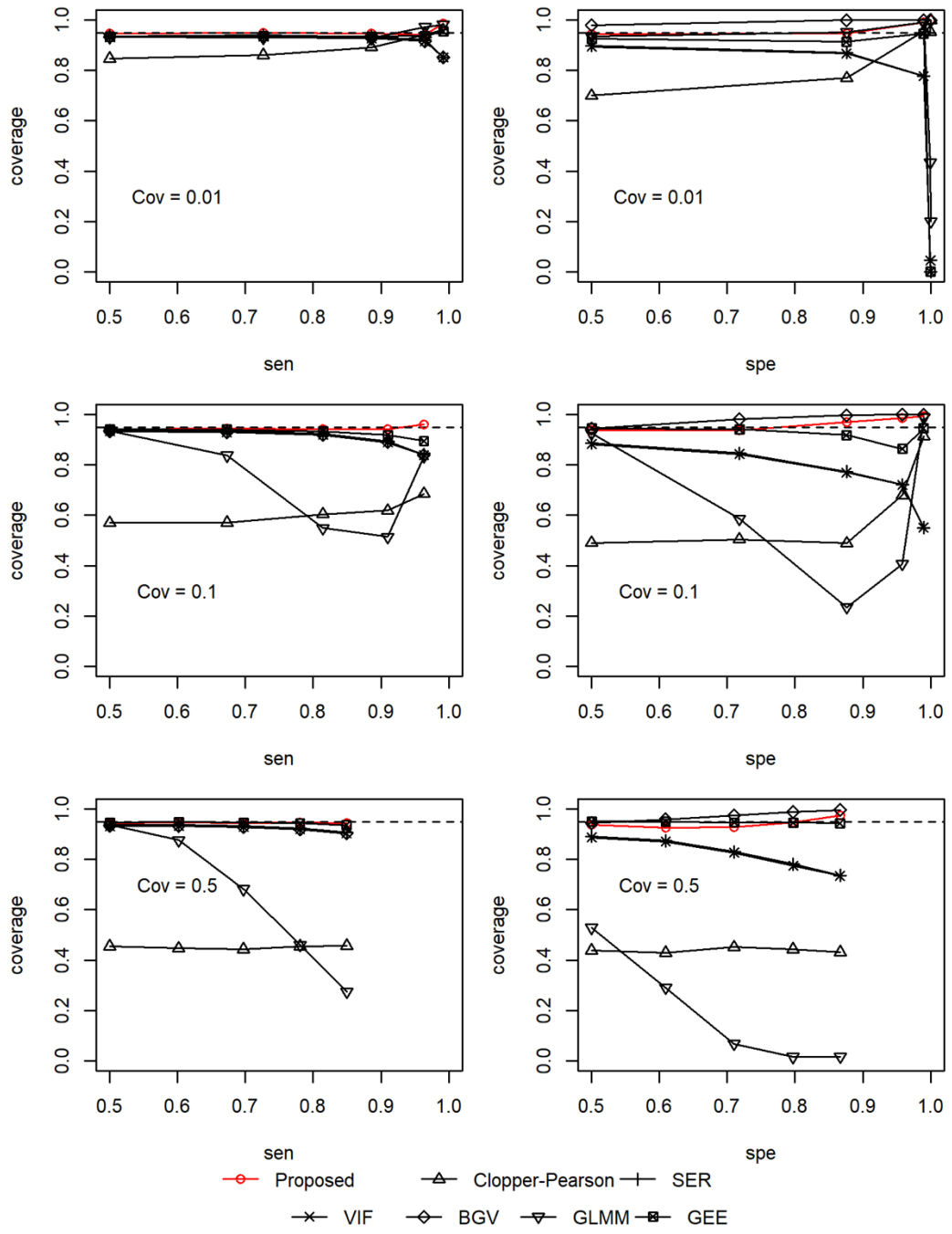

Figure 2.2. Coverage probabilities of 95% CIs of sensitivity and specificity from the simulation study. Solid lines with different symbols denote different methods. The dashed black line is the nominal CP(= 95%).

Table 2.1. The number of converged analyses out of 10,000 simulations for different sensitivity and specificity values (round to 4 decimal places)

Cutoff			−0.8	−0.6	−0.4	−0.2	0
Sensitivity	Cov = 0.01	**True value**	**0.5**	**0.7268**	**0.8861**	**0.9648**	**0.9921**
		GLMM	10000	10000	8831	9968	9688
		GEE	10000	10000	10000	10000	9688
	Cov = 0.1	**True value**	**0.5**	**0.6726**	**0.8145**	**0.9101**	**0.9632**
		GLMM	10000	9817	9016	9752	9496
		GEE	10000	10000	10000	10000	9980
	Cov = 0.5	**True value**	**0.5**	**0.6019**	**0.6972**	**0.7807**	**0.8492**
		GLMM	9557	9530	9448	9438	9542
		GEE	10000	10000	10000	10000	10000
Specificity	Cov = 0.01	**True value**	**1**	**0.9997**	**0.9895**	**0.8759**	**0.5**
		GLMM	5	451	7621	9762	9929
		GEE	5	458	7796	10000	10000
	Cov = 0.1	**True value**	**0.9895**	**0.9584**	**0.8759**	**0.7181**	**0.5**
		GLMM	5365	9188	9483	9350	9217
		GEE	5504	9369	9997	10000	10000
	Cov = 0.5	**True value**	**0.8664**	**0.7973**	**0.7105**	**0.6092**	**0.5**
		GLMM	9624	9633	9601	9403	9293
		GEE	9981	9999	10000	10000	10000

2.4 Application

2.4.1 Data collection

The real data were collected from a study on the influenza A virus (IAV) in swine. Animals in this study were chosen from a commercial farm in which there were about 600 breeding females [Panyasing 2014]. Approximately 21-day-old piglets were weaned, ear-tagged, and quarantined in one room. The piglets were randomly assigned to 6 treatments by assigning ear tags to treatments randomly. The treatments were the combinations of inoculation (None, H1N1 and H3N2) and vaccination (Yes and No), where H1N1 and H3N2 are two

subtypes of IAV. Vaccination was given to the piglets on the day post inoculation (DPI) -42 and -21. On DPI -10, they were shipped to Iowa State University and the piglets with the same treatments were grouped into several pens. On DPI 0, the piglets received either H1N1 or H3N2 inoculation. Oral fluid samples were collected weekly at DPI = 0, 7, 14, 21, 28, 35, and 42 from each pen. Thus samples could be identified to the pen level. Antibodies were measured by NP-blocking-ELISA with 3 technical repeats for each sample. Samples, as well as the positive controls and negative controls, were measured in plates. Positive and negative controls are used for the quality control, which are expected to provide outputs within certain ranges. If either the negative control or the positive control or both provides an output out of the range, the plate is invalid and the samples in the same plate need to be re-tested. The sample to negative (S/N) is the response, which is defined as

$$S/N = \frac{\text{Sample} - \overline{\text{Positive control}}}{\overline{\text{Negative control}} - \overline{\text{Positive control}}}.$$

A separate statistical analysis [Panyasing 2014] showed that 1) vaccination did not make any significant difference; 2) there was no significant difference between the two subtypes of IAV. Therefore, vaccination is excluded from our data analysis and H1N1 and H3N3 are combined as “Yes” for inoculation.

Totally, samples were collected from 26 pens. Piglets from 4 of the 26 pens did not receive any inoculations while piglets from the other 22 pens received either H1N1 or H3N2 inoculation. Totally, there were 3 (repeats) $\times$ 7 (DPI) = 21 observations from each pen. The logarithm of S/N was considered as the response so that the distribution of the response was more symmetric. To estimate the sensitivity and specificity, the population was divided into two groups: the non-diseased group and the diseased group. The observations from the non-diseased group had two sources: 1) samples collected from pens which did not receive any

inoculation; 2) samples collected when DPI = 0 even if the piglets in the pen were inoculated. The rest were classified as from the diseased group. Therefore, the non-diseased samples were collected from 26 pens, 4 of which were sampled 7×3 times while the other 22 were sampled 3 times. In the diseased group, there were 22 pens, each of which were sampled 6×3 times.

2.4.2 Results

The real data was analyzed by the linear mixed model and the point estimates of model parameters were: $\hat{\mu}_0 = -0.06$, $\hat{\mu}_1 = -0.86$, $\hat{\sigma}_\gamma = 0.11$, $\hat{\sigma}_{e_0} = 0.15$ and $\hat{\sigma}_{e_1} = 0.33$. The cutoff values were selected so that the sensitivity or specificity could reach desired values. In this paper, the empirical estimate of specificity was set to be 0.9 and 0.95. The corresponding cutoff values, the point estimates and CI's (keeping 3 decimal digits) of sensitivity and specificity for the real data were calculated and shown in Table 2.2.

2.5 Conclusion

The method proposed in this study provides a new way to calculate the CIs of sensitivity and specificity for correlated and normally distributed data. In the new method, variances of $\widehat{\text{Sen}}$ and $\widehat{\text{Spe}}$ are calculated based on the multi-variate normal distribution. The logit transformation and the corresponding back-transformation are applied to ensure the range of the calculated CIs is limited to [0,1]. From the simulation, the CIs calculated using the proposed method overall have more reasonable CPs (close to 95% if the significance level is 0.05) than the CIs from the current methods. Moreover, the proposed method reduces the risk of the convergence problem and doesn't require re-analysis of data if we choose different cutoff values. This is an important advantage over current binary model-based methods, especially when performing high volumes of simulations with different cutoff values.

Table 2.2. The estimates and 95% CIs of sensitivity and specificity of the real data calculated by the proposed method and the existing method.

cutoff			−0.235	−0.356
Specificity	Empirical Estimate		0.900	0.947
	CI	Proposed	(0.805, 0.952)	(0.848, 0.983)
		Clopper-Pearson	(0.840, 0.943)	(0.898, 0.977)
		Wilson	(0.842, 0.938)	(0.898, 0.973)
		Simple Asymptotic	(0.852, 0.948)	(0.911, 0.983)
		Agresti-Coull	(0.840, 0.939)	(0.896, 0.974)
		SER	(0.808, 0.992)	(0.889, 1.000)
		VIF	(0.810, 0.990)	(0.890, 1.000)
		BGV	(0.444, 1.000)	(0.501, 1.000)
		GLMM	(0.991, 1.000)	(0.977, 1.000)
		GEE	(0.841, 0.939)	(0.897, 0.973)
Sensitivity	Empirical Estimate		0.949	0.919
	CI	Proposed	(0.912, 0.972)	(0.878, 0.947)
		Clopper-Pearson	(0.923, 0.969)	(0.888, 0.944)
		Wilson	(0.923, 0.967)	(0.888, 0.942)
		Simple Asymptotic	(0.928, 0.971)	(0.892, 0.946)
		Agresti-Coull	(0.923, 0.967)	(0.888, 0.942)
		SER	(0.918, 0.981)	(0.880, 0.958)
		VIF	(0.919, 0.980)	(0.881, 0.957)
		BGV	(0.919, 0.980)	(0.881, 0.957)
		GLMM	(0.926, 0.991)	(0.892, 0.969)
		GEE	(0.923, 0.967)	(0.888, 0.942)

CHAPTER 3. A BAYESIAN MODEL FOR CLUSTERED DATA WITH LATENT VARIABLES

Abstract: The evaluation of new diagnostic tests for diseases are challenging, especially when the true statuses are uncertain. For example, an enzyme-linked immunosorbent assay (ELISA) test detects and measures antibodies in blood, but not all diseased subjects will have obvious antibody reaction when exposed to certain antigens. In this study, we proposed a Bayesian hierarchical model for estimation of sensitivities and specificities of 6 new diagnostic tests measured on the same subjects. We use latent variables to indicate the antibody reaction statuses, which are not known for sure. Simulation studies show that our model provides accurate estimates of the model parameters and sensitivities and specificities of the diagnostic tests.

3.1 Introduction

As discussed in Section 1.1, sensitivity and specificity are building bricks of an ROC curve which is an important statistical method to evaluate and compare binary tests. Sensitivity, also known as the true positive rate, is the proportion of positive tests within the diseased group. Similarly, specificity, known as the true negative rate, is the proportion of negative tests within the non-diseased group. If the statuses of subjects are known, the sensitivity and specificity are estimable by the empirical method. However, in many clinical tests, the true statuses of the subjects are unknown. In this situation, the empirical method will fail and the model-based method is the only way to estimate the sensitivity and specificity. Given a cutoff (or threshold) value that separates positive and negative tests, and under the assumption that observations in the diseased group tend to possess larger values, the model-based estimates of sensitivity and specificity are defined in formulas (2.5 – 2.6) in Chapter 2. Estimating the model parameters,

which is required for the model-based estimates of sensitivity and specificity, tends to be challenging because of 1) intra-subject correlation; 2) unknown subject status. For example, in a clinical test, if a person gets infected with some virus, the counts of some of the relevant antibodies might increase to fight the virus. However, the types of antibodies that increase in counts might be different for different patients.

Methods to analyze clustered data have been developed for decades and well described [Galbraith 2010]. [Dunson 2000] incorporated latent variables to analyze mixed outcomes. The latent variables, either assumed to be normally distributed or linked to variables that followed a simple exponential family, were used to control correlations between observations. However, the model parameters are not complicated enough to solve the problems discussed in this paper. In our study, the antibody statuses of observations are unknown and are modeled by a latent variable. We use the Bayesian method combined with Markov chain Monte Carlo (MCMC) to make inference for the model parameters [Gelman 2004, Gilk 1996, and Robert 2004].

3.2 Data

3.2.1 Data Collection

The data were collected in a veterinary clinical study. In this study, 24 14-week old pigs (subjects) were randomly assigned to 4 groups to receive Actinobacillus pleuropneumoniae (APP) with serovars 1, 5, 7 or 12 by blindly selecting ear tags from a bag, with 6 pigs per group. Serum samples were collected weekly and oral fluid samples were collected daily till the 56^{th} day post inoculation (DPI). Antibodies were tested using enzyme-linked immunosorbent assay (ELISA) tests. Totally, seven antibodies ("Serum-S/P_ApxIV_IgG", "Serum-S/P_ApxIV_IgM", "Serum-S/P_ApxIV_IgA", "APP.1(9,11)-ELISA", "5a,5b_LPS_ELISA", "7(4)_LPS_ELISA", and "12_LPS_ELISA") from the serum

samples and three antibodies (“S/P_IgG”, “S/P_IgM”, and “S/P_IgA”) from the oral fluid samples were tested.

3.2.2 Data properties

All the 24 pigs were healthy initially and got infected because of the inoculation that would stimulate some antibody to fight against the virus. Figure 3.1 shows the observations for pig 77 and pig 78 that were both inoculated with the virus with serotype 1.

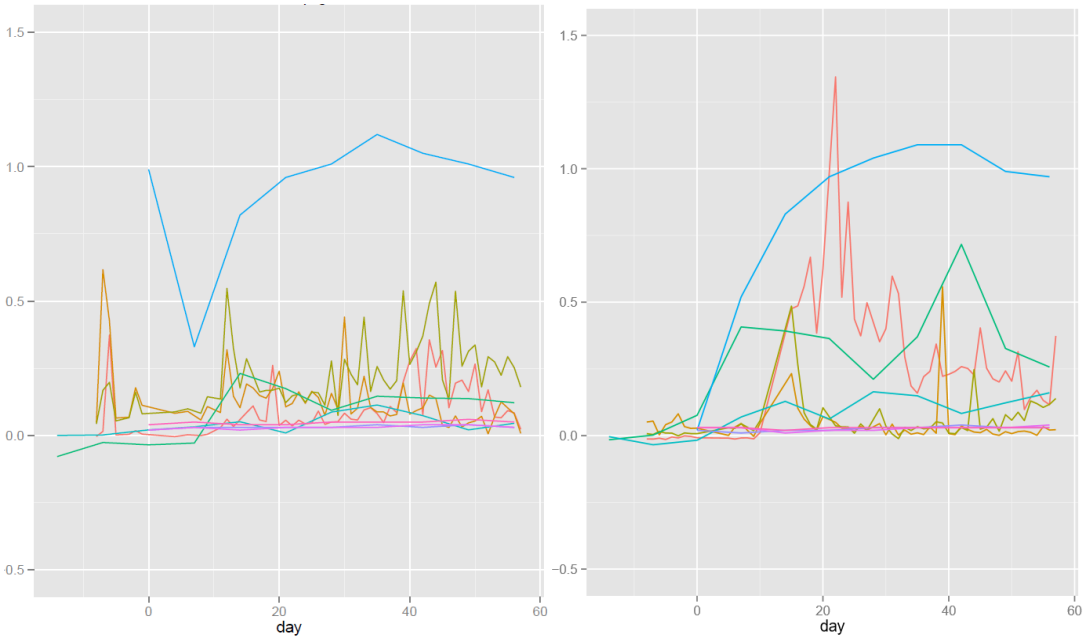

Figure 3.1. Observations of all antibodies for pig 77 (left) and pig 78 (right): different colors denote different antibodies.

From Figure 3.1, we observed the following. 1) All of the antibodies from both pigs stayed at the baseline for day<0, indicating that both pigs were noninfected initially. 2) After the inoculation, in both pigs, the counts of some of the antibodies increased dramatically from the baseline and reached to much higher levels. Here, a new binary variable, named antibody status, was introduced and it could be either normal (at the baseline) or abnormal (at the higher

level). 3) For the infected pigs, antibodies that showed abnormality were random: the antibody denoted by the dark green curve showed abnormality in pig 78, but remained at the normal status in pig 77.

Figure 3.2 shows the observations of the antibody named "S/P_IgG" from all the pigs.

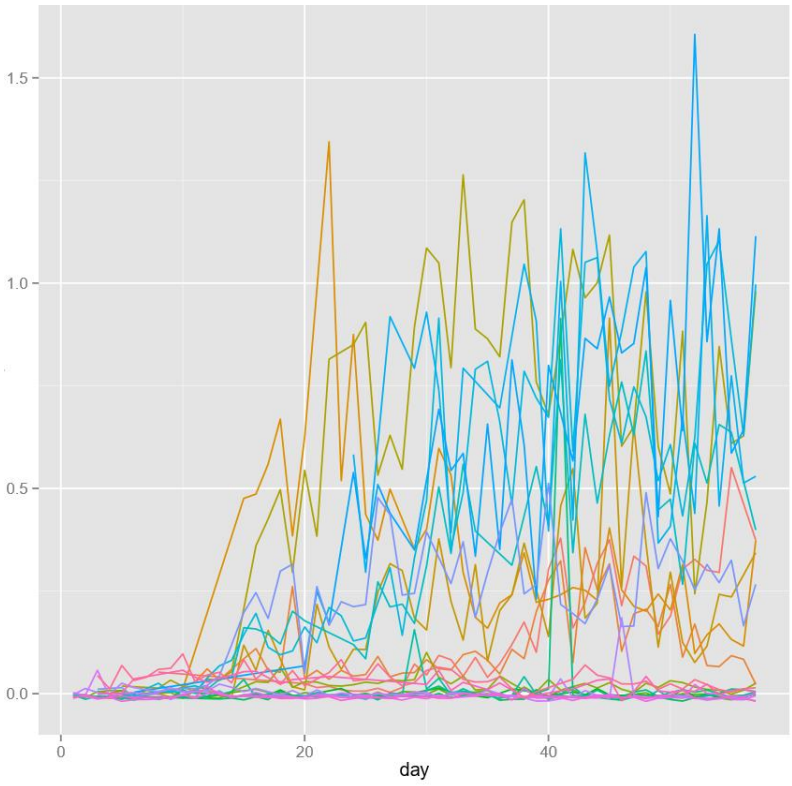

Figure 3.2. Observations of the antibody named "S/P_IgG" from all the pigs: different colors denote different pigs.

From Figure 3.2, we observed that for the antibody, the variances of observations were different for different antibody statuses: the abnormal antibodies, with higher values, possessed larger variances.

This paper proposes a normal-distribution-based hierarchical model with the latent status. We use the Bayesian method together with MCMC to analyze the data.

3.3 Method

3.3.1 Model

During the process of model building, some facts discussed previously were taken into consideration: 1) For the infected pigs, some (not necessarily all) of the antibodies would show abnormality and the antibodies that showed abnormality might be different for different infected pigs. 2) The standard errors were heterogeneous for different antibodies and different antibody statuses. Besides these facts, three assumptions were made for this model: 1) Data (or transformed data) were normally distributed; 2) Observations of the same pig were correlated. 3) Once an antibody reached the abnormal status, the antibody would stay at the abnormal level. In this study, the observations for an antibody given a certain pig might have different statuses at different time points, i.e., initially normal and then abnormal sometime after the inoculation. This would increase the complexity of the model. Because the change of a certain antibody over time was not our point of interest, data were truncated and those with DPI<20 were excluded for analysis. For the truncated data, observations of the same pig and the same antibody roughly stayed at either the normal status or the abnormal status throughout the rest of the time.

We also observed that no antibody was stimulated to fight against the virus for some pigs that were inoculated with the virus with serotype 12, which might be indicative of a different population. In this study, all pigs in the group of serotype 12 were excluded for data analysis.

Let Y_{iljk} denote the measurement of the k^{th} antibody measured from the i^{th} pig that received the l^{th} serotype on the j^{th} day , i =1, 2,…, 6; j=21, 22,…, 56; k=1, 2,…, 10; l =1, 2, and 3. Based on these facts and assumptions, we proposed the following model:

$$Y_{iljk} = \beta_{0k} + \beta_{1k} S_{ilk} + (z_{il} + \varepsilon_{iljk})\sigma_{S,k},$$

where

$$S_{ilk} \overset{\text{indep}}{\sim} \text{Bernoulli}(P_{kl})$$

and β_{0k} is the mean of the k^{th} antibody if it is at the level of normality while β_{1k} is the mean difference between the abnormal level and the normal level for the k^{th} antibody; P_{kl} denotes the probability of the k^{th} antibody to reach the abnormal level for pigs inoculated with the virus of the serotype l; S_{ilk} denotes the status (0 for the normal level and 1 for the abnormal level) of the k^{th} antibody for the i^{th} pig that received virus with serotype l; $z_{il} \overset{\text{iid}}{\sim} N(0, \rho)$ and $\varepsilon_{iljk} \overset{\text{iid}}{\sim} N(0, 1-\rho)$ are random effects to account for intra-subject (pig) correlation and random error, respectively, where ρ denotes the correlation of observations within the same pig; $\sigma_{S,k}$, an unknown constant, denotes the standard deviation (sd) of the k^{th} antibody at level S. Note: in the following context, I will call the i^{th} pig that received the l^{th} serotype as the $(il)^{\text{th}}$ pig.

From this model, we have that

$$\mu_{ilk} \triangleq \beta_{0k} + \beta_{1k} S_{ilk}$$

$$= \begin{cases} \beta_{0k} & \text{if the } k^{\text{th}} \text{antibody of the } (il)^{\text{th}} \text{pig is at the normal status} \\ \beta_{0k} + \beta_{1k} & \text{otherwise} \end{cases}$$

$$\mathrm{cov}\left(Y_{iljk}, Y_{i'l'j'k'} | S_{ilk},\ S_{i'l'k'}\right)$$

$$= \rho\sigma_{Sk}\sigma_{S\prime k\prime} I(i = i') + (1-\rho)\sigma_{Sk}\sigma_{S\prime k\prime} I(i = i')I(j = j')I(k = k')$$

$$= \begin{cases} 0 & \text{for different pigs} \\ \rho\sigma_{Sk}\sigma_{S\prime k\prime} & \text{for the same pig but either different antibodies or different times} \\ \sigma_{Sk}^2 & \text{for the same pig, antibody, and time} \end{cases}$$

Because of the complexity of this hierarchical model, the likelihood function was hard to obtain. Thus, the Bayesian method was used during the process of data analysis. Because we didn't have enough information for model parameters, non-informative priors were assigned [Gelman 2004, Gelman 2006, Huang 2013, and Kerman 2011]:

$$\beta_{0k} \sim N(0, 1000)$$

$$\beta_{1k} \sim N(0, 1000)I(0, +\infty)$$

$$\sigma_{S,k} \sim \mathrm{unif}(0, 100)$$

$$P_{k,l} \sim \mathrm{Beta}(0.5, 0.5)$$

$$\rho \sim \mathrm{unif}(-1, 1)$$

3.3.2 Method to determine the cutoff

Once we have parameter values, the cutoff values are determined by a desired value of either sensitivity or specificity. In this study, the true status is the antibody status. Therefore, sensitivity and specificity can be calculated with respect to the antibody status, and the inversion of the formula for sensitivity or the specificity can be used to calculate the cutoffs. If the kth antibody from the $(il)^{\text{th}}$ pig was at the abnormal status, i.e., S_{ilk}=1, then the observations for this antibody would follow a normal distribution with mean = $\beta_{0k} + \beta_{1k}$ and sd = σ_{1k}, i.e.,

$$Y_{iljk} \sim N(\beta_{0k} + \beta_{1k}, \sigma_{1k}^2).$$

Alternatively, if an antibody was at the normal status, i.e., S_{ilk}=0, then the observations for this antibody would follow a normal distribution with mean = β_{0k} and sd = σ_{0k}, i.e.,

$$Y_{iljk} \sim N(\beta_{0k}, \sigma_{0k}^2).$$

Therefore, for the k^{th} antibody, given a desired value of sensitivity or specificity, denoted by $\widehat{\text{Sen}}_{\text{k}}$ and $\widehat{\text{Spe}}_{\text{k}}$, the cutoff is calculated by one of the two following two methods:

$$\begin{aligned}\widehat{\text{Sen}}_{\text{k}} &= \widehat{\Pr}\left(Y_{iljk} > c_k \middle| S_{ilk} = 1\right) \\ &= \widehat{\Pr}\left(Y_{iljk} - \hat{\beta}_{0k} - \hat{\beta}_{1k} > c_k - \hat{\beta}_{0k} - \hat{\beta}_{1k} \middle| S_{ilk} = 1\right) \\ &= 1 - \Phi\left(\frac{c_k - \hat{\beta}_{0k} - \hat{\beta}_{1k}}{\hat{\sigma}_{1k}}\right).\end{aligned}$$

Therefore,

$$c_k = Q\left(1 - \widehat{\text{Sen}}_{\text{k}}\right) \cdot \hat{\sigma}_{1k} + \hat{\beta}_{0k} + \hat{\beta}_{1k},$$

where $\Phi(.)$ and $Q(.)$ are the cumulative distribution function and quantile function of $N(0, 1)$. Similarly,

$$\begin{aligned}\widehat{\text{Spe}}_{\text{k}} &= \widehat{\Pr}\left(Y_{iljk} < c_k \middle| S_{ilk} = 0\right) \\ &= \widehat{\Pr}\left(Y_{iljk} - \hat{\beta}_{0k} < c_k - \hat{\beta}_{0k} \middle| S_{ilk} = 0\right) \\ &= \Phi\left(\frac{c_{\text{k}} - \hat{\beta}_{0k}}{\hat{\sigma}_{0k}}\right).\end{aligned}$$

And

$$c_k = Q\left(\widehat{\text{Spe}}_{\text{k}}\right) \cdot \hat{\sigma}_{0k} + \hat{\beta}_{0k}$$

3.4 Results

3.4.1 Convergence

Since some antibodies stayed at the normal status all the time for all the pigs and couldn't provide enough information about the difference between the two antibody statuses, only six antibodies ("S/P_IgG", "S/P_IgM", "S/P_IgA", "APP.1(9,11)-ELISA", "5a,5b_LPS_ELISA", and "7(4)_LPS_ELISA") were chosen for data analysis. The data were analyzed with R: rjags. Two Markov chains were used for the MCMC algorithm. First, 100,000 iterations were used for the burn-in process. Then another 100,000 iterations were used for sampling, which reflected the posterior distributions of the parameters. The trace-plots of the 100,000 samples from the posterior distributions of some key parameters were made. Figure 3.3 shows the trace-plots of the sampled β_1's only.

Visual inspection of the trace-plots showed that the sampled β_1's from the two Markov chains roughly converge to the same distributions. The Gelman-Rubin diagnostic [Brooks 1998 and Gelman 1992] was applied to the samples of all key parameters. The Gelman-Rubin diagnostic is mainly used to test the convergence of multiple Markov chains. For univariate diagnostic, the variance within chains, denoted by *W*, and the variance between chains, denoted by *B*, are calculated first. The overall variance, denoted by *V*, is the weighted average of *B* and *W*. It assesses whether the difference between *V* and *W* is large enough to be a concern. The Gelman-Rubin statistic, called the potential scale reduction factor (PSRF), is defined as

$$\hat{R} = \sqrt{\frac{V}{W}}.$$

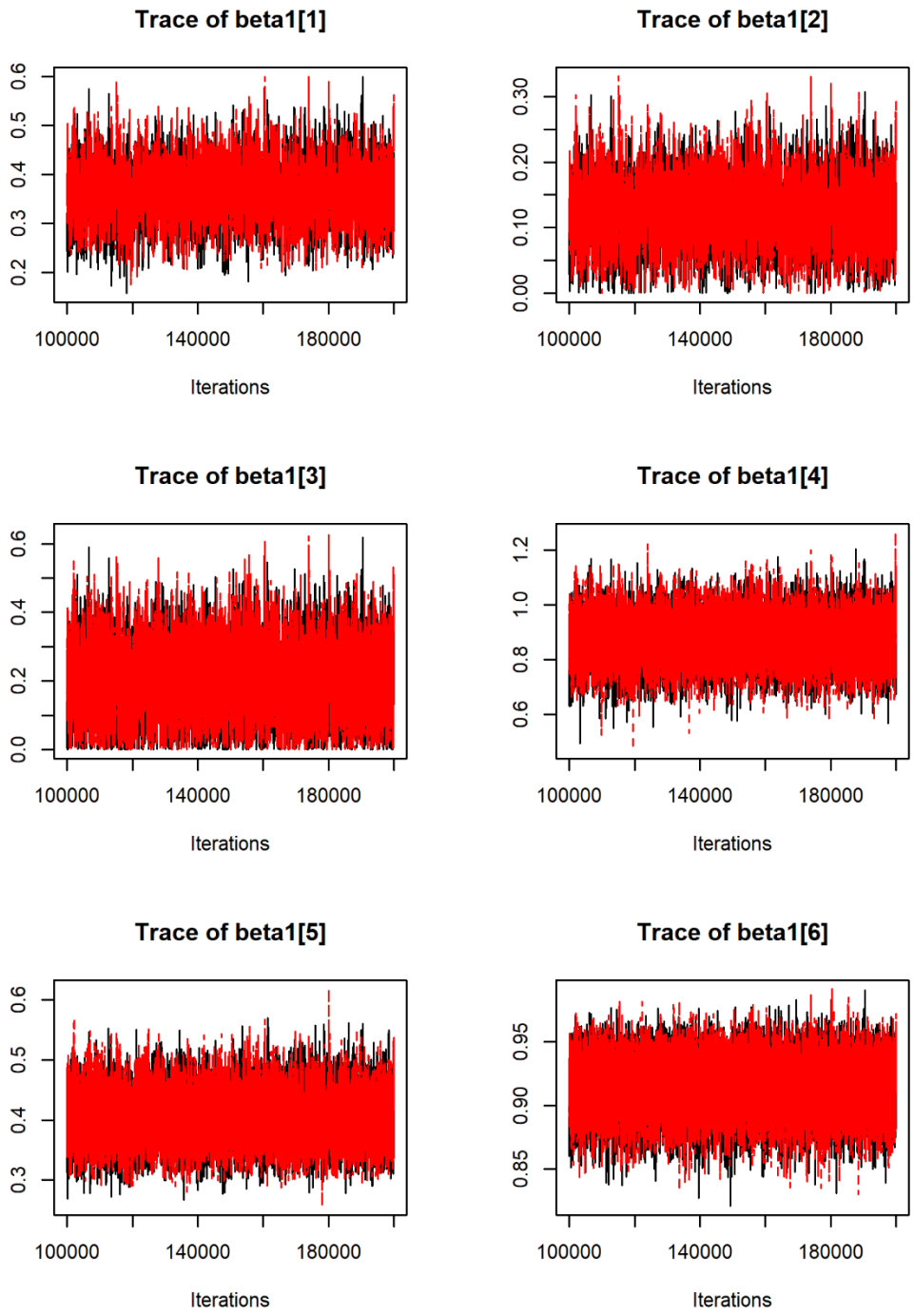

Figure 3.3. Trace-plots of the sampled β_1's

If $\hat{R}$ is larger than 1.1, it is a concern. From the univariate test, the maximum of the upper limit of CIs of *R*'s is 1.01. The PSRF of the multivariate Gelman-Rubin diagnostic [Brooks 1998] turned out to be 1. Therefore, based on either the univariate or the multivariate Gelman-Rubin diagnostic, the MCMC sampling process converged.

3.4.2 Estimates of model parameters

The model parameters were estimated by the sample means and credible intervals (c.i.'s) obtained from the posterior distributions. Table 3.1 shows the estimates of the key parameters (means, variances, and correlation coefficient) that were useful for evaluating sensitivity and specificity or justifying assumptions when establishing the model.

Table 3.1. Point estimates and 95% c.i.'s of the key model parameters.

Parameter	β_{01}	β_{02}	β_{03}	β_{04}	β_{05}	β_{06}	β_{11}	β_{12}	β_{13}
Estimate	0.035	0.052	0.142	0.051	0.036	0.04	0.363	0.123	0.208
c.i.	0.016 ~ 0.056	0.033 ~ 0.072	0.109 ~ 0.178	0.039 ~ 0.064	0.031 ~ 0.041	0.032 ~ 0.048	0.269 ~ 0.463	0.046 ~ 0.206	0.045 ~ 0.379
Parameter	β_{14}	β_{15}	β_{16}	σ_{01}	σ_{02}	σ_{03}	σ_{04}	σ_{05}	σ_{06}
Estimate	0.877	0.401	0.915	0.062	0.062	0.111	0.035	0.014	0.023
c.i.	0.747 ~ 1.007	0.339 ~ 0.471	0.882 ~ 0.946	0.052 ~ 0.085	0.051 ~ 0.085	0.094 ~ 0.151	0.027 ~ 0.049	0.011 ~ 0.02	0.018 ~ 0.033
Parameter	σ_{11}	σ_{12}	σ_{13}	σ_{14}	σ_{15}	σ_{16}	ρ	P_{11}	P_{21}
Estimate	0.37	0.306	0.607	0.332	0.168	0.086	0.383	0.75	0.372
c.i.	0.312 ~ 0.502	0.254 ~ 0.417	0.5 ~ 0.828	0.242 ~ 0.483	0.125 ~ 0.241	0.062 ~ 0.125	0.197 ~ 0.676	0.372 ~ 0.977	0.051 ~ 0.77

Table 3.1 continued

Parameter	P_{31}	P_{41}	P_{51}	P_{61}	P_{12}	P_{22}	P_{32}	P_{42}	P_{52}
Estimate	0.25	0.916	0.084	0.083	0.214	0.499	0.071	0.071	0.929
c.i.	0.022 ~ 0.63	0.62 ~ 1	0 ~ 0.381	0 ~ 0.379	0.019 ~ 0.557	0.166 ~ 0.833	0 ~ 0.331	0 ~ 0.329	0.671 ~ 1
Parameter	P_{62}	P_{13}	P_{23}	P_{33}	P_{43}	P_{53}	P_{63}		
Estimate	0.072	0.917	0.417	0.75	0.084	0.071	0.928		
c.i.	0 ~ 0.334	0.622 ~ 1	0.094 ~ 0.79	0.372 ~ 0.978	0 ~ 0.381	0 ~ 0.329	0.669 ~ 1		

From Table 3.1, we observed the following. 1) For all antibodies, the variances at the abnormal status were ~5-10 times as large as those at the normal status, which indicated that the heterogeneous variance assumption for different antibody statuses was reasonable. 2) The estimated correlation coefficient $\hat{\rho} = 0.38$ and the c.i.$= (0.2, 0.68)$ confirmed the existence of the correlation between observations from the same subject. A follow-up question was that: "within the same subject, is it possible that observations from the same antibody are more correlated than those from different antibodies?" That was possible. However, given the limited number of subjects, it was not worth introducing heterogeneous correlations for different antibodies. Thus, this estimated correlation could be viewed as an averaged correlation for all antibodies within the same subject. We also observed that the 4^{th} – 6^{th} serum antibodies, "APP.1(9,11)-ELISA", "5a,5b_LPS_ELISA", and "7(4)_LPS_ELISA", were significantly responsive to infections corresponding to serovars 1, 5, and 7, respectively with

$P_{j+3,j} > 90\%, j = 1, 2, \text{and } 3$. However, the cross responses of serum antibodies to serovars were very low with $P_{ij} < 10\%, j = 1, 2, \text{and } 3; i \geq 4 \text{ and } i \neq i + 3$. The responses of oral fluid antibodies to the different infections were also quite different with most stimulation rates varying between 21.4% and 75%, and very few (=2) rates <10% or >90%.

Cutoff values were calculated from each of the sampled parameter values by following the discussion in Section 3.3 and thus could be considered "sampled" cutoffs. The point estimates and c.i.'s of the cutoffs w.r.t $\widehat{sen} = 90\%$ were calculated by the means and the 2.5% - 97.5% percentiles of the "sampled" cutoffs and were shown in Table 3.2. An alternative way for the point estimates of cutoffs is to calculate the cutoffs based on the estimated model parameters. The point estimates of cutoffs calculated by these two methods turned out to be identical if we kept two decimal digits.

Table 3.2. Point estimates and 95% c.i.'s of the cutoffs for the six antibodies.

Parameter	c_1	c_2	c_3	c_4	c_5	c_6
Estimate	-0.076	-0.216	-0.43	0.502	0.221	0.845
c.i.	-0.292 - 0.054	-0.396 - -0.107	-0.794 - -0.193	0.244 - 0.676	0.108 - 0.296	0.774 - 0.893

Figure 3.4 shows the distributions of the six antibodies as well as the cutoff values determined with respect to a 90% sensitivity. From Figure 3.4, we observed that the abnormal data and the normal data were clearly separated for the serum antibodies ("APP.1(9,11)-ELISA", "5a,5b_LPS_ELISA", and "7(4)_LPS_ELISA"), which made it possible to achieve large sensitivity and specificity simultaneously. Furthermore, the majority of the data at the normal level were distributed below the cutoff values, which led to specificity close to 1. On the contrary, for the oral fluid antibodies, the differences between the normal data and the

abnormal data were shaded by the comparably-sized variances, making it hard to obtain large sensitivity and specificity simultaneously. Therefore, from the perspective of clinical testing, antibodies from the serum samples were better choices.

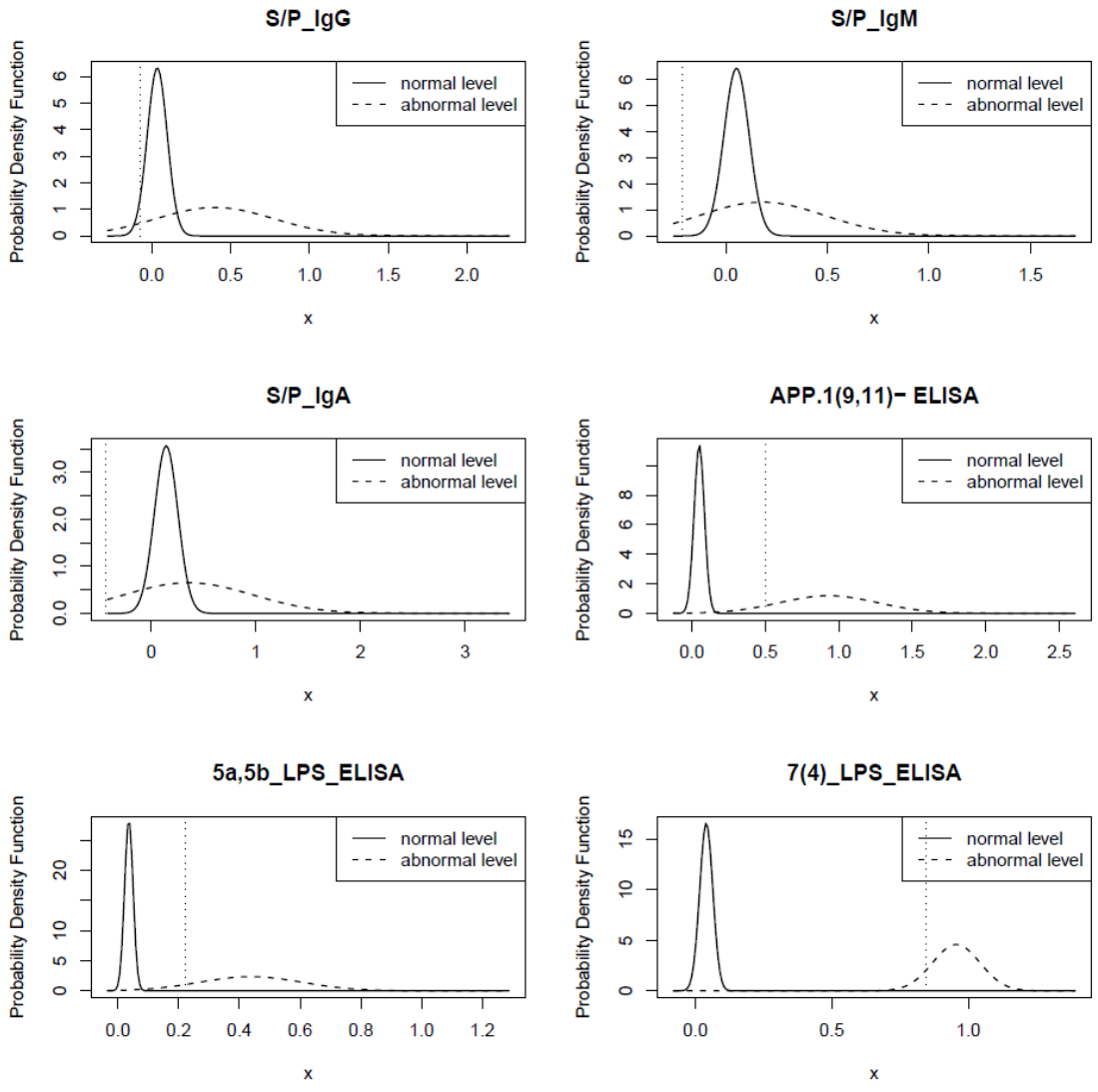

Figure 3.4. Plots of the probability density function (PDF) for the six antibodies: the solid black lines denote the PDF for the normal level; the dashed lines denote the PDF for the abnormal level; the dotted lines denote the determined cutoff values to achieve 90% sensitivity.

3.5 Simulations

Accurately estimated parameters are important for estimating the distributions of response variables and thus the cutoff values for given values of sensitivity or specificity. To evaluate how accurate the proposed method can recover the true parameters, we applied this method to simulated data. There were six response variables in the simulated data. The first three were sampled "daily" and there were 35 sampled values on each subject for these antibodies. The other three response variables were sampled "weekly" and had 5 values on each subject. These six response variables were mimics of antibodies from the oral fluid and serum samples. The parameters for simulations were chosen to be in a similar range of the estimated parameters from the real data and are shown in Table 3.3. Note that β_0's were all chosen to be 0 since they don't influence the relative magnitudes between sensitivity and specificity.

Table 3.3. True parameters for simulations

Parameter	β_{01}	β_{02}	β_{03}	β_{04}	β_{05}	β_{06}	β_{11}	β_{12}
True value	0	0	0	0	0	0	0.15	0.3
Parameter	β_{13}	β_{14}	β_{15}	β_{16}	σ_{01}	σ_{02}	σ_{03}	σ_{04}
True value	0.45	0.6	0.75	0.9	0.05	0.05	0.05	0.05
Parameter	σ_{05}	σ_{06}	σ_{11}	σ_{12}	σ_{13}	σ_{14}	σ_{15}	σ_{16}
True value	0.05	0.05	0.3	0.3	0.3	0.3	0.3	0.3
Parameter	ρ		P_{kl} $k \leq 3, l \leq 3$		P_{kl} $6 \geq k > 3,$ $l = k - 3$		P_{kl} $6 \geq k > 3,$ $l \neq k - 3$ & $l \leq 3$	
True value	0.4		0.5		0.9		0.1	

The number of subjects (pigs) was chosen to be either 18 (3 serotypes × 6 subjects/serotype) or 180 (3 serotypes × 60 subjects/serotype). For each sample size, 200 datasets were simulated. The proposed method was applied to the simulated data with 30,000 iterations for both the burn-in process and the sampling process. Cutoffs were calculated with respect to each of the 30,000 samples of parameters for sensitivity values of 0.5, 0.6, 0.7, 0.8, and 0.9. The point estimates and c.i.'s of model parameters and cutoffs were obtained from each simulation. The averages, biases, and standard errors of these estimates as well as the coverage probabilities of c.i.'s of key parameters ($\beta's$ and $\sigma's$) and cutoffs are shown in Tables 3.4 and 3.5, respectively.

Table 3.4. Averages, biases and standard errors of the point estimates as well as the c.i.'s coverage probabilities of the key parameters

	180 pigs				TRUE value	18 pigs			
	Avg	Bias	SE	Coverage		Avg	Bias	SE	Coverage
β_{01}	0	0	0.002	0.955	0	−0.002	−0.002	0.013	0.95
β_{02}	0	0	0.002	0.97	0	−0.001	−0.001	0.01	0.955
β_{03}	0	0	0.002	0.97	0	−0.001	−0.001	0.01	0.955
β_{04}	0	0	0.003	0.98	0	−0.001	−0.001	0.011	0.96
β_{05}	0	0	0.003	0.94	0	0	0	0.011	0.945
β_{06}	0	0	0.003	0.96	0	−0.001	−0.001	0.011	0.955
β_{11}	0.15	0	0.012	0.975	0.15	0.147	−0.003	0.043	0.965
β_{12}	0.301	0.001	0.012	0.97	0.3	0.298	−0.002	0.047	0.95
β_{13}	0.45	0	0.012	0.97	0.45	0.446	−0.004	0.048	0.95
β_{14}	0.6	0	0.018	0.945	0.6	0.591	−0.009	0.064	0.96
β_{15}	0.751	0.001	0.017	0.965	0.75	0.745	−0.005	0.066	0.95

Table 3.4 continued

β_{16}	0.9	0	0.016	0.95	0.9	0.896	−0.004	0.06	0.955
σ_{01}	0.05	0	0.002	0.96	0.05	0.07	0.02	0.071	0.815
σ_{11}	0.303	0.003	0.009	0.935	0.3	0.344	0.044	0.076	0.78
σ_{02}	0.05	0	0.002	0.91	0.05	0.059	0.009	0.012	0.795
σ_{12}	0.302	0.002	0.009	0.955	0.3	0.355	0.055	0.071	0.805
σ_{03}	0.051	0.001	0.003	0.96	0.05	0.059	0.009	0.012	0.79
σ_{13}	0.303	0.003	0.009	0.935	0.3	0.356	0.056	0.071	0.805
σ_{04}	0.051	0.001	0.003	0.955	0.05	0.06	0.01	0.013	0.85
σ_{14}	0.303	0.003	0.013	0.96	0.3	0.367	0.067	0.091	0.84
σ_{05}	0.05	0	0.002	0.915	0.05	0.06	0.01	0.014	0.84
σ_{15}	0.302	0.002	0.013	0.945	0.3	0.367	0.067	0.09	0.86
σ_{06}	0.05	0	0.002	0.955	0.05	0.06	0.01	0.013	0.87
σ_{16}	0.302	0.002	0.013	0.94	0.3	0.36	0.06	0.083	0.89

Table 3.5. Averages, biases and standard errors of the point estimates as well as the c.i.'s coverage probabilities of the cutoffs

Sensitivity		180 pigs				TRUE value	18 pigs			
		Avg	Bias	SE	Coverage		Avg	Bias	SE	Coverage
0.5	c_1	0.15	0	0.014	0.98	0.15	0.145	−0.005	0.055	0.955
	c_2	0.301	0.001	0.014	0.975	0.3	0.297	−0.003	0.056	0.95
	c_3	0.45	0	0.014	0.975	0.45	0.445	−0.005	0.056	0.955
	c_4	0.601	0.001	0.019	0.95	0.6	0.591	−0.009	0.071	0.955
	c_5	0.751	0.001	0.018	0.97	0.75	0.745	−0.005	0.073	0.945
	c_6	0.901	0.001	0.018	0.95	0.9	0.895	−0.005	0.067	0.96

Table 3.5 continued

0.6	c_1	0.074	0	0.014	0.975	0.074	0.058	−0.016	0.052	0.96
	c_2	0.224	0	0.014	0.965	0.224	0.207	−0.017	0.063	0.95
	c_3	0.373	−0.001	0.014	0.96	0.374	0.355	−0.019	0.063	0.95
	c_4	0.524	0	0.02	0.935	0.524	0.498	−0.026	0.08	0.955
	c_5	0.675	0.001	0.019	0.97	0.674	0.652	−0.022	0.08	0.945
	c_6	0.824	0	0.018	0.95	0.824	0.804	−0.02	0.073	0.955
0.7	c_1	-0.008	−0.001	0.014	0.975	−0.007	−0.035	−0.028	0.057	0.955
	c_2	0.142	−0.001	0.014	0.965	0.143	0.111	−0.032	0.074	0.945
	c_3	0.291	−0.001	0.014	0.965	0.293	0.258	−0.034	0.074	0.945
	c_4	0.442	−0.001	0.021	0.93	0.443	0.398	−0.044	0.096	0.94
	c_5	0.593	0	0.02	0.96	0.593	0.552	−0.041	0.093	0.93
	c_6	0.742	−0.001	0.019	0.95	0.743	0.706	−0.036	0.086	0.945
0.8	c_1	−0.104	−0.002	0.015	0.965	−0.102	−0.144	−0.042	0.07	0.975
	c_2	0.046	−0.001	0.015	0.965	0.048	−0.002	−0.05	0.091	0.93
	c_3	0.195	−0.002	0.016	0.955	0.198	0.145	−0.052	0.091	0.925
	c_4	0.346	−0.002	0.023	0.93	0.348	0.282	−0.066	0.118	0.92
	c_5	0.497	0	0.022	0.95	0.498	0.436	−0.062	0.113	0.92
	c_6	0.646	−0.002	0.021	0.95	0.648	0.592	−0.055	0.105	0.93
0.9	c_1	−0.237	−0.003	0.017	0.965	−0.234	−0.295	−0.061	0.095	0.92
	c_2	−0.087	−0.002	0.017	0.96	−0.084	−0.158	−0.074	0.117	0.9
	c_3	0.062	−0.003	0.018	0.955	0.066	−0.011	−0.077	0.117	0.89
	c_4	0.213	−0.003	0.027	0.95	0.216	0.121	−0.095	0.152	0.91
	c_5	0.365	−0.001	0.025	0.94	0.366	0.274	−0.091	0.146	0.885
	c_6	0.513	−0.003	0.025	0.945	0.516	0.434	−0.082	0.135	0.9

From Tables 3.4 and 3.5, we observed the following. 1) The averaged estimates were very close to the true values for the simulation studies for 180 subjects (pigs). The averaged estimates from 18 subjects were significantly deviated from the true values, especially for variances and cutoffs with extreme values. 2) The standard errors of the estimates of key parameters and cutoffs were smaller for simulations with 180 subjects. 3) The coverage probabilities of the 95% c.i.'s, 91% ~ 98%, were close to the value of 95% for simulations with 180 subjects. For simulations with 18 subjects, the coverage probabilities related to group means and cutoffs ($\leq$0.7) were close to 95%. However, the coverage probabilities for variances and cutoffs ($\geq$0.8) were far below 95%.

3.6 Conclusion

This paper proposed a model based on the normal distribution to analyze clustered data with latent classes. The Bayesian method together with MCMC was used because of the complexity of this hierarchical model. The data structure in the example of this paper is prevalent nowadays, especially in clinical tests. However, it is difficult to analyze such data because 1) intra-subject correlations exist if multiple observations were obtained from each of the subjects; 2) during the process of calculating the sensitivity and specificity, the true status is usually unknown, which makes the model-based estimates the only choice. The true status was included in the proposed model as a latent variable. From the simulation study, the proposed method can be used to analyze this kind of data and make accurate inferences to the true parameters if the sample size is large enough (~180). The estimates of variances and cutoffs associated with extreme sensitivity (or specificity) values might be problematic if the sample size is as small as 18.

CHAPTER 4. R PACKAGE: CICorr

Purpose: As discussed previously, many clinical diagnostic tests involve multiple measurements from the same subjects, generating intra-subject correlated data values. The calculations of the CIs of sensitivity and specificity become challenging because of this correlation. In Chapter2, we propose a new method to solve this problem, which outperforms current methods according to the coverage probabilities of CIs and the calculating efficiency. I create an R package for the 1st project (Chapter 2) so that everyone can install and use it. The manual of the R package is attached.

4.1 Basic information

Package name CICorr

Version 0.3

Date 2017-09-25

Title calculate the confidence intervals of sensitivity and specificity for intra-subject correlated and normally distributed data

Author Y. Du and C. Wang

Maintainer Y. Du

Description This package is used to calculate the confidence intervals of sensitivity and specificity for normally-distributed data with intra-subject correlations given cutoff values.

Depends R (>= 3.4.4), nlme, mnormt

4.2 Contents of manual

CIcorr-package	*The 'CICorr' package: summary information*

Description:

This package provides a function for computing the CIs of sensitivity and specificity for intro-subject correlated and normally distributed data w.r.t given cutoffs.

Details:

The package includes a function 'CI_corr' which is used to calculate the CI of sensitivity and specificity for intra-subject correlated and normally distributed data.

License:

Author(s):

Y. Du (R programmer and package creator), C. Wang (Advisor of methodology).

References:

Du Y., 2018. Choosing cutoff values for correlated continuous diagnostic data to estimate sensitivity and specificity (Doctoral dissertation).

CI_corr	*Function to calculate CI of sensitivity and specificity*

Description:

The function is used to calculate the CIs of sensitivity and specificity for a correlated and normally distributed data w.r.t given cutoff values.

Usage:

```
CI_corr(subject, status, y, cut, method)
```

Arguments:

subject:	a character vector of length d, where d is the number of observations, representing the subject ID for each observation.
status:	a vector of either 0 or 1 with length d, representing the true status: 0=healthy and 1=ill
y:	a numerical vector of length d, representing observations
cut:	specified numerical cutoff(s)
method:	1 or 2. If 1, the point estimates of sensitivity and specificity in the CIs are selected from the model-based estimate and empirical estimate depending on different situation. If 2, the Agresti-Coull estimates are used.

Details:

This function analyzes data with the linear mixed model to obtain estimates of model parameters. The estimates of sensitivity and specificity were calculated for given cutoffs. Variance of Empirical sensitivity and specificity were calculated which were functions of sensitivity and specificity, respectively. Logit transformation and the back-transformation was applied to the calculated sensitivity and specificity to make sure the final CIs were between [0,1]. The point estimates of sensitivity and specificity in the CIs can be either a selection among empirical estimation and model-based estimation or estimates based on the Agresti-Coull method.

Values:

Returns a matrix including the point estimates, upper confidence limits and lower confidence limits of sensitivity and specificity w.r.t specified cutoff(s).

Examples:

Compute the CIs of sensitivity and specificity for data "clinical" which possesses intra-subject correlations.

```
data(clinical)
CI_corr(subject=clinical$pen,status=(clinical$status=="ill"),y=clinical$log,cut=seq(-
1,0,length.out=4),method=2)
```

4.3 R code

```
#' CI_corr
#'
#' Take in subject, status, response and cutoff and then calculate CI for sensitivity and
specificity
#'
#' @name CI_corr
#'
#' @param subject: a character vector of length d, where d is the number of observations,
representing the subject ID for each observation.
#' @param status: a vector of either 0 or 1 with length d, representing the true status: 0=healthy
and 1=ill
#' @param y: a numerical vector of length d, representing observations
#' @param cut: specified numerical cutoff(s)
```

#' @param method: 1 or 2, the point estimate in the CI is the mixture of model based estimate and empirical estimate for "1" and Agresti-Coull estimate for "2"

#' @return a matrix includes low and high limits of CI of sensitivity and specificity for different cutoffs

#' @seealso \code{lme}

#' @seealso \code{pmnorm}

#'

#' @details

#' This function runs the linear mixed regression for y with the random subject effect to obtain estimates of model parameters.

#' Empirical, model-based and Agresti-Coull estimates of sensitivity and specificity are calculated for given cutoffs. Variance of Empirical

#' sensitivity and specificity are calculated which are functions of sensitivity and specificity, respectively. Logit transformation and the

#' back-transformation are applied to the calculated sensitivity and specificity to make sure the final CIs are between [0,1].

#' @importFrom mnormt pmnorm

#' @importFrom nlme lme fixef VarCorr

#'

#' @examples

#' ## Compute the CIs of sensitivity and specificity for data "clinical" which possesses intra-cluster correlations.

#' data(clinical)

```
#'       CI_corr(subject=clinical$pen,status=(clinical$status=="ill"),y=clinical$log,cut=c(-0.5,-
0.6),method=2)
#'
#' @export
CI_corr=function(subject,status,y,cut,method) {
 CI = matrix(, length(cut), 6)
   colnames(CI) = c("Spe.est", "Spe.L", "Spe,H", "Sen.est",
      "Sen.L", "Sen.H")
   delta = 5
   data = data.frame(subject = as.factor(subject), status = as.factor(status),
      y)
   data = data[order(data$status), ]
   x = data$y
   fit = nlme::lme(fixed = y ~ status, random = list(subject = pdSymm(form = ~1)),
      weights = varIdent(form = ~1 | status), data = data)
   m.hat = as.vector(nlme::fixef(fit))
   sd <- as.numeric(nlme::VarCorr(fit)[, 2])
   V = sd^2
   se.coef = c(1, as.numeric(coef(fit$modelStruct$varStruct,
      unconstrained = F)))
   mse.coef = se.coef^2
   sigma.hat <- matrix(c(mse.coef[1] * V[2], 0, 0, mse.coef[2] *
      V[2]), 2, 2) + V[1]
   sigma0.hat <- matrix(c(mse.coef[1] * V[2], 0, 0, mse.coef[1] *
```

```
    V[2]), 2, 2) + V[1]
sigma1.hat <- matrix(c(mse.coef[2] * V[2], 0, 0, mse.coef[2] *
    V[2]), 2, 2) + V[1]
uni_status = unique(as.vector(status))
uni_status = uni_status[order(uni_status)]
n0 = as.numeric(table(data[data$status == uni_status[1],
    1]))
N0 = sum(n0)
n1 = as.numeric(table(data[data$status == uni_status[2],
    1]))
N1 = sum(n1)
if (m.hat[2] > 0) {
    sen.m <- 1 - pnorm((cut - sum(m.hat))/sqrt(sigma.hat[2,
        2]))
    spe.m <- pnorm((cut - m.hat[1])/sqrt(sigma.hat[1, 1]))
}
else {
    sen.m <- pnorm((cut - sum(m.hat))/sqrt(sigma.hat[2, 2]))
    spe.m <- 1 - pnorm((cut - m.hat[1])/sqrt(sigma.hat[1,
        1]))
}
for (i in 1:length(cut)) {
    if (m.hat[2] > 0) {
```

```
    sen.hat <- sum(x[(N0 + 1):(N0 + N1)] >= cut[i])/N1
    spe.hat <- sum(x[1:N0] < cut[i])/N0
    sen.ac <- (sum(x[(N0 + 1):(N0 + N1)] >= cut[i]) +
       2)/(N1 + 4)
    spe.ac <- (sum(x[1:N0] < cut[i]) + 2)/(N0 + 4)
    spe_var.hat <- (spe.m[i])/N0 - (spe.m[i])^2 * (N0 +
       sum(n0 * (n0 - 1)))/N0^2 + sum(n0 * (n0 - 1))/N0^2 *
       mnormt::pmnorm(rep((cut[i] - m.hat[1]), 2), varcov = sigma0.hat)
    sen_var.hat <- (1 - sen.m[i])/N1 - (1 - sen.m[i])^2 *
       (N1 + sum(n1 * (n1 - 1)))/N1^2 + sum(n1 * (n1 -
       1))/N1^2 * mnormt::pmnorm(rep((cut[i] - sum(m.hat)),
       2), varcov = sigma1.hat)
}
else {
    sen.hat <- sum(x[(N0 + 1):(N0 + N1)] <= cut[i])/N1
    sen.ac <- (sum(x[(N0 + 1):(N0 + N1)] <= cut[i]) +
       2)/(N1 + 4)
    spe.hat <- sum(x[1:N0] > cut[i])/N0
    spe.ac <- (sum(x[1:N0] > cut[i]) + 2)/(N0 + 4)
    spe_var.hat <- (1 - spe.m[i])/N0 - (1 - spe.m[i])^2 *
       (N0 + sum(n0 * (n0 - 1)))/N0^2 + sum(n0 * (n0 -
       1))/N0^2 * mnormt::pmnorm(rep((cut[i] - m.hat[1]),
       2), varcov = sigma0.hat)
```

```
    sen_var.hat <- (sen.m[i])/N1 - (sen.m[i])^2 * (N1 +
        sum(n1 * (n1 - 1)))/N1^2 + sum(n1 * (n1 - 1))/N1^2 *
        mnormt::pmnorm(rep((cut[i] - sum(m.hat)), 2),
          varcov = sigma1.hat)
}
if (method == 1) {
    sen.final = ifelse(sen.hat > 1 - delta/N1 | sen.hat <
        delta/N1, sen.m[i], sen.hat)
    spe.final = ifelse(spe.hat > 1 - delta/N0 | spe.hat <
        delta/N0, spe.m[i], spe.hat)
}
else {
    sen.final = sen.ac
    spe.final = spe.ac
}
M.var <- sen_var.hat/sen.m[i]^2/(1 - sen.m[i])^2
M.est <- log(sen.final/(1 - sen.final))
M.l <- M.est - 1.96 * sqrt(M.var)
M.h <- M.est + 1.96 * sqrt(M.var)
CI[i, 4:6] = exp(c(M.est, M.l, M.h))/(1 + exp(c(M.est,
    M.l, M.h)))
M.var <- spe_var.hat/spe.m[i]^2/(1 - spe.m[i])^2
M.est <- log(spe.final/(1 - spe.final))
```

```
        M.l <- M.est - 1.96 * sqrt(M.var)
        M.h <- M.est + 1.96 * sqrt(M.var)
        CI[i, 1:3] = exp(c(M.est, M.l, M.h))/(1 + exp(c(M.est,
            M.l, M.h)))
    }
  return(CI)
}
```

CHAPTER 5. SUMMARY AND FUTURE WORK

5.1 Summary

In this dissertation, we propose appropriate statistical methods focusing on two challenges when working on the diagnostic test data: correlations and unknown statuses. Chapter 2 develops a method to calculate the CIs of sensitivity and specificity for correlated and continuous data with given true statuses. In chapter 3, we propose a hierarchical Bayesian model to analyze the data with unknown statuses and further to estimate the sensitivity and specificity. We also create an R package for the 1st project for future use. Both methods are developed for clinical diagnostic test data, but their applications are not limited to diagnostic tests. They apply to projects of other areas if the data properties are similar.

In Chapter 2, we propose a new method to calculate the CIs of sensitivity and specificity for intra-subject-correlated data based on the normal distribution. In the new method, variances of $\widehat{\text{Sen}}$ and $\widehat{\text{Spe}}$ are calculated as a function of model parameters. The logit transformation and the corresponding back-transformation are used to ensure the ranges of the calculated CIs are limited to [0,1]. The simulation studies show that the CIs calculated using the proposed method overall possess more reasonable CPs and reduced the risk of the convergence problem. In addition, the new method is more efficient than the current binary-data-based methods because different choices of cutoff values don't require re-analysis of data. This is an important advantage when multiple cutoff values are required. We created an R package for this method for future use.

In Chapter 3, we propose a model based on the normal distribution to analyze clustered data with unknown statuses. The unknown statuses are modeled by a latent variable. The

Bayesian method is used because of the complexity of this hierarchical model. From the simulation study, the proposed method can be used to analyze this kind of data and make accurate inferences to the true parameters if the sample size is large enough (180) and thus can be used to select cutoff values given borderline values of sensitivity or specificity.

5.2 Future work

The first method is based on the normal distribution. We can obviously extend the main idea to the data following a distribution other than the normal distribution. However, the derivation of variances of the estimated sensitivity and specificity might be a big challenge.

In the second project, the model only includes one latent variable because it is reasonable to assume all subjects are in the status of "infection" and we only need to model the unknown antibody status. However, in a diagnostic test, it is not a rare case if both the subject status and the antibody status are unknown. We can obviously extend our idea and model the two statuses with two latent variables. The challenge point is that the probability that an antibody has a reaction is dependent on the disease status of the subject. If there is no further assumptions to simplify the model, the model will include 50+ parameters, which makes the convergence hard if not impossible. Then how to simplify the model? We can try assuming all the 4 serovars have similar influences and thus the probabilities to stimulate a type of antibody don't differ from the types of serovars. Although this assumption might not reasonable in biology, it is worth trying from the perspective of statistics. Finally, if we encounter both correlated data and unknown statuses in a project, can we find an effective way to calculate the CIs of sensitivity and specificity? All these questions need further investigation.

REFERENCES

Agresti A. and Coull B.A., 1998. Approximate is better than "exact" for interval estimation of binomial proportions. The American Statistician 52, 119-126.

Brooks S.P. and Gelman A., 1998. General methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics 4, 434-455.

Brown L.D., Cai T., and DasGupta A., 2001. Interval estimation for a binomial proportion. Statistical Science 16, 101-133.

Clopper C.J. and Pearson E.S., 1934. The use of confidence or fiducial limits illustrated in the case of binomial. Biometrika 26, 404-413.

DeLong E.R., DeLong D.M., and Clarke-Pearson D.L., 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44, 837-845.

Dunson D.B., 2000. Bayesian latent variable models for clustered mixed outcomes. Journal of the Royal Statistical Society: Series B 62, 355-366.

Fleiss J., Levin B., and Paik M.C., 2003. Statistical methods for rates and proportions. Hoboken, NJ: Wiley.

Galbraith S., Daniel J.A., and Vissel B., 2010. A study of clustered data and approaches to its analysis. The Journal of Neuroscience 30, 10601-10608.

Gelman A. and Donald R., 1992. Inference from iterative simulation using multiple sequences. Statistical Science 7, 457-472.

Gelman A., Carlin J., Stern H., and Rubin D., 2004. Bayesian Data Analysis. 2nd edition. Chapman and Hall.

Gelman A., 2006. Prior distributions for variance parameters in hierarchical models. Bayesian Analysis 1, 515-533.

Gilks W., Richardson S., and Spiegelhalter D., 1996. Markov Chain Monte Carlo in Practice. 1st edition. Chapman and Hall.

Hanley J.A. and McNeil B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143, 29-36.

Huang A. and Wand M.P., 2013. Simple marginally noninformative prior distributions for covariance matrices. Bayesian Analysis 8, 439-452.

Kerman J., 2011. Neutral noninformative and informative conjugate beta and gamma prior distribution. Electronic Journal of Statistics 5, 1450-1470.

Metz C.E., 1978. Basic principles of ROC analysis. Seminars in Nuclear Medicine 8, 283-298.

Mutsvari T., Lesaffre E., García-Zattera M., Diya L., and Declerck D., 2010. Factors that influence data quality in caries experience detection: a multilevel modeling approach. Caries Research 44, 438–444.

Newcombe R.G., 1998. Two-sided confidence intervals for the single proportion: comparison of seven methods. Statistics in Medince 17, 857-872.

Panyasing Y., Goodell C.K., Kittawornrat A., Wang C., Kittawornrat A., Prickett J.R., Schwartz K.J., Ballagi A., Lizano S., and Zimmerman J.J., 2014. Detection of influenza A virus nucleoprotein antibodies in oral fluid specimens from pigs infected under experimental conditions using a blocking ELISA. Transboundary and Emerging Disease 61,177-184.

Rao J. and Scott A., 1992. A simple method for the analysis of clustered binary data. Biometrics 48, 577–585.

Robert C. and Casella G., 2004. Monte Carlo Statistical Methods. 2nd edition.
Springer.

Smith P. and Hadgu A., 1992. Sensitivity and specificity for correlated observations. Statistics in Medicine 11, 1503–1509.

Williams R., 2000. A note on robust variance estimation for cluster-correlated data. Biometrics 56, 645–646.

Wilson E.B., 1927. Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association 22, 209-212.

Zeger S. and Liang K., 1986. Longitudinal data analysis for discrete and continuous outcomes. Biometrics 42, 121–130.